(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2003 (18.09.2003)

PCT

(10) International Publication Number
WO 03/076944 A1

(51) International Patent Classification[7]: G01N 33/68, C07K 1/28, C12Q 1/37

(21) International Application Number: PCT/AU03/00300

(22) International Filing Date: 13 March 2003 (13.03.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PS 1118 13 March 2002 (13.03.2002) AU

(71) Applicant *(for all designated States except US)*: **PROTEOME SYSTEMS INTELLECTUAL PROPERTY PTY LTD** [AU/AU]; Unit 1, 35 - 41 Waterloo Road, North Ryde, NSW 2113 (AU).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **WILKINS, Marc** [US/AU]; 118 Annandale Street, Annandale, NSW 2038 (AU). **ARTHUR, Jonathan, Wesley** [AU/AU]; 123 Balaka Drive, Carlingford, NSW 2118 (AU).

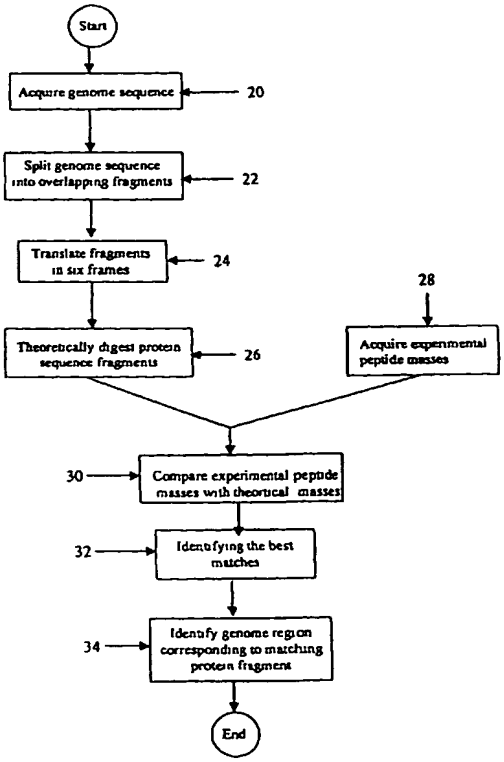
(74) Agent: **F B RICE & CO**; 605 Darling Street, Balmain, NSW 2041 (AU).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: ANNOTATION OF GENOME SEQUENCES



(57) Abstract: A method of identifying one or more proteins in an unannotated DNA sequence is disclosed. The method involves dividing the DNA sequence into a plurality of sequence fragments of substantially the same length (about 300 to 5000 base pairs, most typically 1000 to 1050 base pairs. A six frame translation is then performed on each of the DNA sequence fragments to obtain six translated amino acid sequence fragments for each DNA sequence fragment. Each of the translated sequence fragments is subjected to theoretical digestion to obtain a plurality of cleaved peptide sequences. Next experimental empirical data for peptide fragments from a protein digested in the same manner as the theoretical digestion is compared with the theoretical data generated in step for each of the translated sequence fragments to identify one or more translated sequence fragments which include a substantial number of peptides present in the digested protein. The sequence fragment which has the greatest number of theoretical peptide masses correlating to the empirical data indicates the likely location of the protein of interest in the DNA sequence. To avoid problem where the sequence is divided at the site of a protein, the DNA sequence is duplicated and the original and duplicate are split in such a manner that the sequence fragments from the original overlap the cuts in the original genome sequence.

WO 03/076944 A1